



TITLE:

# 非定常な推移確率分布をもつマルコフ決定過程 (動的計画法の研究会報告集)

AUTHOR(S):

古川, 長太

---

CITATION:

古川, 長太. 非定常な推移確率分布をもつマルコフ決定過程 (動的計画法の研究会報告集). 数理解析研究所講究録 1969, 67: 1-12

ISSUE DATE:

1969-02

URL:

<http://hdl.handle.net/2433/107898>

RIGHT:

## 非定常な推移確率分布をもつ

## マルコフ決定過程

九州大学 理学部 古川 長 太

## §1 序

マルコフ決定過程の殆んどすべての論文に見られるように、  
ここでは、推移確率分布が定常であることに基礎を置く。こ  
の推移確率分布の定常性は、割引率の有る場合では、最適政  
策に関するいくつかのエレガントな性質と深く本質的な素因  
となっている。例えば、Blackwell [2] は、一般に  $(p, \gamma)$  最適  
政策の存在と、特に action が可付番個の場合は  $\gamma$  最適政策、  
action が有限個の場合は最適政策の存在を証明し、かつ、  
これらがいづれも定常政策として得られることを示した。ま  
た、Strauch [3] は、Blackwell より強い意味での  $(p, \gamma)$  最  
適定常政策の存在を証明した。

この報告では、推移確率分布の定常性をゆるめ、非定常な  
推移確率分布をもつものとして、いくつかの意味での最適政  
策の存在と、存在すれば、その最適政策の性質について考察

するにとしよう。

## §2 諸定義 (I)

### 定義 2.1

$X, Y$ ; ある separable metric space の Borel subset

$\mathcal{C}$ ;  $X$  の subset の Borel 集合族

$\mathcal{D}$ ;  $Y$  の subset の Borel 集合族

$P(x)$ ;  $X$  の上の probability distribution の class

$g(y|x)$ ; conditional probability distribution

(各  $x$  につき,  $g(\cdot|x)$  は  $Y$  上の prob. distribution

各  $B \in \mathcal{D}$  につき,  $g(B|\cdot)$  は  $X$  上の  $\mathcal{C}$  可測関数)

$Q(Y|X)$ ;  $g(y|x)$  の class

$M(X)$ ;  $X$  上の 有界実数値可測関数の class

### 定義 2.2

定義 2.1 を拡張して

$P(x_1, x_2, \dots, x_n), P(x_1, x_2, \dots)$ ,

$Q(x_{n+1}|x_1, x_2, \dots, x_n), Q(x_{n+1}, x_{n+2}|x_1, x_2, \dots, x_n)$

$M(x_1, x_2, \dots, x_n)$  等を定義する。

### 定義 2.3

$pu$ ;  $pu \equiv \int_X u(x) dP(x)$  for  $P \in P(x), u \in M(x)$

$gu$ ;  $gu \equiv \int_Y u(x, y) dg(y|x)$  for  $g \in Q(Y|x), u \in M(XY)$

定義 2.4

$p \in P(X)$  なる  $p$  が degenerate  $\Leftrightarrow p\{x\} = 1$  for some  $x \in X$

$g \in Q(Y|X)$  なる  $g$  が degenerate  $\Leftrightarrow g(\cdot|x)$  が各  $x$  について degenerate

神題 2.1 (Blackwell [2]) 各  $g \in Q(Y|X)$ , 各  $u \in M(XY)$  に

対して

$$f u \geq g u \quad \text{for all } x \in X$$

なるような degenerate  $f \in Q(Y|X)$  が存在する。

## §3 諸定義 (II)

定義 3.1

$S$ ; state space (non-empty Borel set)

$A$ ; action space (non-empty Borel set)

$S \ni s$ ; state

$A \ni a$ ; action

$Q(S|SA) \ni g_i$ ; 時刻  $i$  における state から 時刻  $i+1$  における state への transition prob. distribution

$M(SAS) \ni r$ ; reward function

$0 < \beta < 1$ ; discount factor

定義 3.2

$H_n \equiv SAS \cdots AS$  ( $2n-1$  factors)

$$\pi_n \in Q(A|H_n), (n=1,2,\dots) \subset \Gamma$$

$$\pi \equiv \{\pi_1, \pi_2, \pi_3, \dots\}; \text{ strategy}$$

### 定義 3.3

random Markov strategy;  $\pi = \{\pi_1, \pi_2, \dots\}$  におい  $\pi_n \in Q(A|S)$   
( $n=1,2,\dots$ )

Markov strategy; random Markov strategy  $\pi = \{\pi_1, \pi_2, \dots\}$  におい  
て, 各  $\pi_n$  が degenerate

Markov strategy と  $\{f_1, f_2, \dots\}$  を書く。

### 定義 3.4

$$e_\pi \equiv \pi_1 g_1, \pi_2 g_2, \dots \in Q(X|S) \text{ に対し } (X = ASAS \dots)$$

$$I_n(\pi, g, v) \equiv e_\pi \left[ \sum_{j=1}^n \beta^{j-1} r(s_j, a_j, s_{j+1}) + \beta^n v \right]$$

$$I_n(\pi, g) \equiv I_n(\pi, g, 0)$$

$$I(\pi, g) \equiv e_\pi \sum_{j=1}^{\infty} \beta^{j-1} r(s_j, a_j, s_{j+1})$$

$Q^*$ ;  $g_n \in Q(S|SA)$  for  $n=1,2,\dots$  なる  $\{g_1, g_2, \dots\}$  の  
class

補題 3.1  $I_n(\pi, g, v) \rightarrow I(\pi, g) \quad (n \rightarrow \infty) \text{ for } v \in M(S),$   
for  $g \in Q^*$

### 定義 3.5

$\pi^*$  が  $(p, \varepsilon, g)$ -optimal;  $p\{I(\pi^*, g) \geq I(\pi, g) - \varepsilon\} = 1$  for all  $\pi$

$\pi^*$  が  $(\varepsilon, g)$ -optimal;  $I(\pi^*, g) \geq I(\pi, g) - \varepsilon$  for all  $\pi$

$\pi^*$  が  $g$ -optimal;  $I(\pi^*, g) \geq I(\pi, g)$  for all  $\pi$

## §4 最適政策

補題 4.1 各  $p \in P(S)$ , 各  $\varepsilon > 0$ , 各  $g \in Q^*$  に対して,  $(p, \varepsilon, g)$ -optimal Markov strategy が存在する。

(証明)

Blackwell [2] の Theorem 2 において  $\{g, g, \dots\}$  の代りに  $\{g_1, g_2, g_3, \dots\}$  とおけば, 直ちに証明される。

定義 4.1

$$T_{nj}; T_{nj}u(s) = \int [r(s, f_n(s), t) + \beta u(t)] d g_j(t | s, f_n(s))$$

$T_{nj}$  を  $(f_n, g_j)$  に対応する operator と呼び,  $(f_n, g_j) \leadsto T_{nj}$  と書く。

$\pi = \{f_1, f_2, \dots\}$  に対して

$$U_j; U_j u(s) = \sup_n T_{nj} u(s) \quad \text{for } (f_n, g_j) \leadsto T_{nj}$$

$U_j$  を  $(\pi, g_j)$  に対応する operator と呼び,  $(\pi, g_j) \leadsto U_j$  と書く。

定理 4.1 (a)  $T_{nj}$  は単調

(b)  $T_{nj}(u+c) = T_{nj}u + \beta c$  for  $c$  is constant

(c) Markov  $\pi = \{f_1, f_2, \dots\}$  に対して,  $(f_n, g_n) \leadsto T_{nn}$  とすれば

$$T_{11} T_{22} \dots T_{nn} v = I_n(\pi, g, v) \quad \text{for each } n$$

(d) Markov  $\pi = \{f_1, f_2, \dots\}$  に対して  $(f_n, g_n) \leadsto T_{nn}$  とすれば

$$T_{nn} I(n\pi, ng) = I(n^{-1}\pi, n^{-1}g) \quad \text{for each } n$$

ただし  $n\pi = \{\pi_{n+1}, \pi_{n+2}, \dots\}, \quad ng = \{g_{n+1}, g_{n+2}, \dots\}$

定義 4.2

$\pi$  を Markov strategy とし、

$f$  が  $\pi$ -generated ;  $f$  は  $S \rightarrow A$  の可測関数で、 $S$  の分割  $\{S_n\}$  が存在して

$$f = f_n \text{ on } S_n \text{ for each } n.$$

$F(\pi)$  ;  $\pi$ -generated function の class

Markov  $\pi' = \{g_1, g_2, \dots\}$  が  $\pi$ -generated ; 各  $g_n$  が  $g_n \in F(\pi)$ .

$G(\pi)$  ;  $\pi$ -generated strategy の class

定理 4.2 (a)  $\pi$  を任意の Markov strategy とする。各  $g_j \in Q(S|SA)$ , 各  $\hat{f} \in F(\pi)$  に対して

$$\hat{T}_j u \leq U_j u$$

ただし  $(\hat{f}, g_j) \rightarrow \hat{T}_j$ 。

(b) 各 Markov  $\pi$ , 各  $\varepsilon > 0$ , 各  $g_j \in Q(S|SA)$  に対して  $\hat{f}_j \in F(\pi)$  が存在し、 $(\hat{f}_j, g_j) \rightarrow \hat{T}_{jj}$  とすれば

$$\hat{T}_{jj} u \geq U_j u - \varepsilon$$

定義 4.3

$(\pi, g_j) \rightarrow U_j$  とし、 $\lim_{n \rightarrow \infty} U_j U_{j+1} \cdots U_n u \Rightarrow u_{j-1}^*$  とおき、これを  $(\pi, g_j)$  に対応する limit point と云う。記号で  $(\pi, g_j) \rightarrow u_{j-1}^*$  と書く。特に  $u_0^* = u^*$  と書く。

定理 4.3 (a)  $\pi$  を任意の Markov strategy とすると

$$I(\hat{\pi}, g) \leq u^* \text{ for } \hat{\pi} \in G(\pi)$$

ただし  $(\pi, g) \rightarrow u^*$ .

(b) 各 Markov  $\pi$ , 各  $\varepsilon > 0$ , 各  $g_j \in Q(S|SA)$  に対して  $\hat{\pi} \in G(\pi)$  が存在して  $I(\hat{\pi}, g) \geq u^* - \varepsilon$ . ただし  $(\pi, g) \rightarrow u^*$ .

(c) もし, 各  $j \geq 0$  に対して  $(\varepsilon, g_j)$ -optimal strategy が存在すれば,  $(\varepsilon/(1-\beta), g)$ -optimal Markov strategy が存在する. ( $\varepsilon \geq 0$ )

(d)  $(f \equiv a, g_j) \rightarrow T_{aj} u_j$  とする. もしも, 各  $\varepsilon > 0$ , 各  $j \geq 0$  に対して  $(\varepsilon, g_j)$ -optimal strategy が存在すれば, Markov strategy  $\hat{\pi}$  が存在して  $(\hat{\pi}, g) \rightarrow \hat{u}_j^*$  については,  $\hat{u}_j^*$  は可測関数でかつ,

$$\hat{u}_{j-1}^* = \sup_{a \in A} T_{aj} \hat{u}_j^* \quad \text{for } j=1, 2, \dots$$

が成立する.

(e) 各  $j \geq 0$  につき,  $\hat{\pi}^*$  が  $g_j$ -optimal  $\iff$

$$I(\hat{\pi}^*, g_j) = \sup_{a \in A} T_{aj} I(\hat{\pi}^*, g) \quad \text{for } j=1, 2, \dots$$

定理 4.3 (c) の系 各  $j \geq 0$  につき  $g_j$ -optimal strategy が存在すれば  $g$ -optimal Markov strategy が存在する.

#### 定義 4.4

action  $a$  と action  $b$  が  $(s, g_j)$  において同等;

$$T_{aj} u(s) = T_{bj} u(s) \quad \text{for all } u \in M(S)$$

action  $a$  と action  $b$  が  $(s, g)$  において同等;

$$T_{aj} u(s) = T_{bj} u(s) \quad \text{for all } u \in M(S),$$

$$\text{for all } g_j \text{ in } \{g_1, g_2, \dots\}$$



$\pi$  を Markov strategy として,

$A$  が essentially countable by  $\pi$ ; 各  $(s, a)$  に対して

$f_n(s)$  と  $a$  が  $(s, g)$  において同値となるような  $n$  が存在する。

$A$  が essentially finite by  $\pi$ ;  $S$  の分割  $\{S_n\}$  が存在して

各  $n$  について,  $s \in S_n$  なる各  $(s, a)$  に対して

$f_1(s), f_2(s), \dots, f_m(s)$  の内の少なくとも一つと  $a$  が  $(s, g)$  において同値となる。

補題 4.2  $A$  が essentially finite by  $\pi$  ならば, 各  $g_j \in Q(S|SA)$ , 各  $u \in M(S)$  に対して  $\hat{f}_j \in F(\pi)$  が存在して,  
 $\hat{T}_{jj} u = \bigcup_j u$  が成立つ。ただし  $(\hat{f}_j, g_j) \rightsquigarrow \hat{T}_{jj}$ ,  
 $(\pi, g_j) \rightsquigarrow \bigcup_j$ 。

定理 4.4 (a)  $A$  が essentially countable by  $\pi$  ならば, 各  $\varepsilon > 0$ , 各  $g \in Q^*$  に対して  $(\varepsilon, g)$ -optimal Markov strategy が存在する。

(b)  $A$  が essentially finite by  $\pi$  ならば, 各  $g \in Q^*$  に対して,  $g$ -optimal strategy が存在する。

## §5 l-stationary strategy.

### 定義 5.1

$g$  が l-stationary;  $g_j \in Q(S|SA)$  ( $j=1, 2, \dots, l$ ) なる

$g_1, g_2, \dots, g_l$  により,  $g = \{\bar{g}, \bar{g}, \bar{g}, \dots\}$  と書けること。ただし  $\bar{g} = \{g_1, g_2, \dots, g_l\}$ 。

$\pi$  が  $l$ -stationary ; degenerate measurable  $f_j: S \rightarrow A$  :

( $j=1, 2, \dots, l$ ) により,  $\pi = \{\bar{f}, \bar{f}, \bar{f}, \dots\}$  と書けること。ただし  $\bar{f} = \{f_1, f_2, \dots, f_l\}$ 。

### 定義 5.2

$\bar{g} = \{g_1, g_2, \dots, g_l\}$ ,  $(\pi, g_j) \rightarrow U_j$  として

$\bar{U}_l \equiv U_1, U_2, \dots, U_l$  ;  $(\pi, \bar{g})$  に対応する operator

記号で  $(\pi, \bar{g}) \rightarrow \bar{U}_l$  と書く。

定理 5.1  $(\pi, \bar{g}) \rightarrow \bar{U}_l$  とすると,  $\bar{U}_l$  は contraction coefficient  $\beta^l$  をもつ contraction mapping である。

定理 5.2 (a)  $\pi$  を Markov strategy,  $g = \bar{g}^{(\infty)}$  が  $l$ -stationary とする。  $(\pi, \bar{g}) \rightarrow \bar{U}_l$  とする。  $\bar{U}_l$  の fixed point を  $\bar{u}_l^*$  とすれば  $I(\hat{\pi}, \bar{g}^{(\infty)}) \leq \bar{u}_l^*$  for all  $\hat{\pi} \in G(\pi)$ 。

(b) 各 Markov  $\pi$ , 各  $l$ -stationary  $g = \bar{g}^{(\infty)}$  に対して,  $\bar{u}_l^*$  を

(a) の枠に定義すれば, 各  $\varepsilon > 0$  に対して  $l$ -stationary strategy

$\bar{f}^{(\infty)} \in G(\pi)$  が存在して,  $I(\bar{f}^{(\infty)}, \bar{g}^{(\infty)}) \geq \bar{u}_l^* - \varepsilon$  が成立つ。

(c) 各  $p \in P(S)$ , 各  $\varepsilon > 0$ , 各  $l$ -stationary  $g = \bar{g}^{(\infty)}$  に対して,

$(p, \varepsilon, \bar{g}^{(\infty)})$ -optimal  $l$ -stationary strategy が存在する。

(d) 各  $\varepsilon \geq 0$  に対して, 各  $j=1, 2, \dots, l$  に対して

$(\varepsilon, (g_j, g_{j+1}, \dots, g_l, \bar{g}^{(\infty)}))$ -optimal strategy が存在すれば,

$(\varepsilon/(1-\beta), \bar{g}^{(0)})$ -optimal  $l$ -stationary strategy が存在する。

(e)  $g$  を  $l$ -stationary とする。このとき

各  $j=1, 2, \dots, l$  に対して  $\delta\pi^*$  が  $\delta g$ -optimal

$$\iff I(\delta^{-1}\pi^*, \delta^{-1}g) = \sup_{a \in A} T_{aj} I(\delta\pi^*, \delta g) \quad (j=1, 2, \dots, l)$$

## §6 Strong optimality.

### 定義 6.1

$\pi^*$  が strong  $(p, \varepsilon, g)$ -optimal ;  $p \{ I(\pi^*, g) \geq \sup_{\pi} I(\pi, g) - \varepsilon \} = 1$

定理 6.1 各  $p \in P(S)$ , 各  $g \in Q^*$ , 各  $\pi$  に対して,

$$p I(\hat{\pi}, g) \geq p I(\pi, g)$$

なるような Markov strategy  $\hat{\pi}$  が存在する。

定理 6.2 各  $p \in P(S)$ , 各  $\varepsilon > 0$ , 各  $g \in Q^*$  に対して,  $(p, \varepsilon, g)$ -optimal strategy が存在する。

定理 6.3 Markov strategy の任意の sequence  $\{\pi^j, j=1, 2, \dots\}$ , 各  $\varepsilon > 0$ , 各  $g \in Q^*$  に対して

$$I(\hat{\pi}, g) \geq \sup_j I(\pi^j, g) - \varepsilon$$

なる Markov  $\hat{\pi}$  が存在する。

定理 6.4 各  $p \in P(S)$ , 各  $\varepsilon > 0$ , 各  $g \in Q^*$  に対して, strong  $(p, \varepsilon, g)$ -optimal Markov strategy が存在する。

注意 Strauch [3] は 特に  $\gamma$  が stationary の場合に, この section の定理 6.4 に相当する結果を, "conservation" の概念を導入することにより与えている。而もその証明方法は, conservation に関するいくつかの補題を用意して, 可成り遠廻りなものである。我々の場合には,  $\gamma$  が stationary でないため, conservation の補題が成立たず, 従って Strauch の方法では成功しないことが分る。然し乍ら, 我々の定理 4.3 (a), (b) を用いることにより, もっと直接的に定理 6.4 を証明することに成功した。

#### 附記

この報告では, 補題, 定理の証明を記載しなかったが, それらの証明については, 参考文献 [4] に詳細に記されている。

#### 参 考 文 献

- [1] R. Bellman, Dynamic Programming, Princeton Univ. Press, (1957).
- [2] D. Blackwell, Discounted dynamic programming, Ann. Math. Stat. 36 (1965), 226-235.
- [3] R.E. Strauch, Negative dynamic programming, Ann. Math. Stat. 37 (1966), 871-890.
- [4] N. Furukawa, A Markov decision process with non-stationary

transition laws, *Bulletin of Mathematical Statistics*,  
13 (1968), 41-52.